BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problems Mailbox.

Japanese Patent Laid-open Publication No. HEI 6-215036 A

Publication date : August 5, 1994

Applicant : XEROX CORPORATION (US)

Title : METHOD FOR SEARCHING FOR DOCUMENT COLLECTION

5

(57) [ABSTRACT]

[OBJECT]

An object of the present invention is to search for a document set which meets a user's requirement from a document collection without the user inputting any searching word.

[STRUCTURE]

A document collection is divided into document sets using a dividing or clustering algorithm (Step 11). A summary is created for each document set using an automatic summarizing algorithm (Step 13). A user can select one or more summaries (Step 15), and a new document collection is formed by a document set corresponding to the selected summary (Step 20). Processings for these division, summary creation and selection are repeated until a document set which meets a user's requirement is found out.

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平6－215036

(43)公開日　平成６年(1994)８月５日

(51)Int.Cl.[5]	識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｇ０６Ｆ　15/40	５００　Ｑ	7218－5Ｌ		
15/401		7218－5Ｌ		

審査請求　未請求　請求項の数１　ＯＬ　（全　４　頁）

(21)出願番号　特願平5－162989

(22)出願日　平成５年(1993)６月30日

(31)優先権主張番号　９８８５３４

(32)優先日　1992年12月10日

(33)優先権主張国　米国（ＵＳ）

(71)出願人　590000798
ゼロックス　コーポレイション
ＸＥＲＯＸ　ＣＯＲＰＯＲＡＴＩＯＮ
アメリカ合衆国　ニューヨーク州　14644
ロチェスター　ゼロックス　スクエア
（番地なし）

(72)発明者　リチャード　ディー．ヘンダーソン
アメリカ合衆国　カリフォルニア州
95128　サン　ホゼ　アーリータ　アベニュー　505

(74)代理人　弁理士　中島　淳　（外２名）

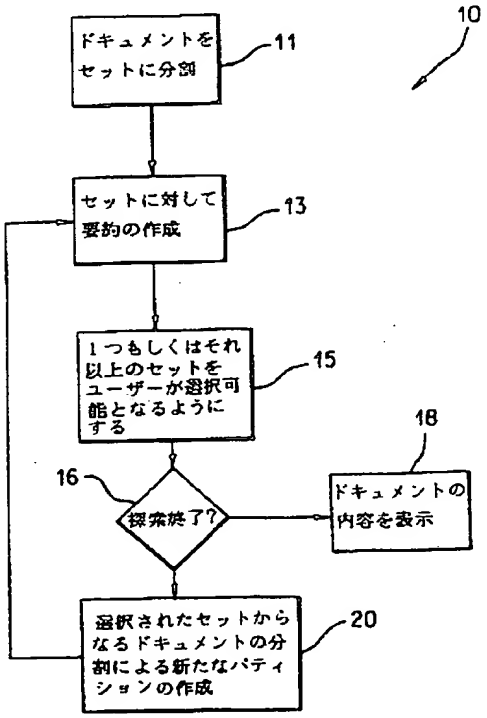
最終頁に続く

(54)【発明の名称】　ドキュメントコレクションの探索方法

(57)【要約】

【目的】　ユーザーが探索ワードを入力せずに、ドキュメントコレクションの中から、ユーザーにとって満足のいくドキュメントセットを探索する。

【構成】　分割あるいはクラスタ化のアルゴリズムを用いてドキュメントコレクションをドキュメントセットに分割する（ステップ１１）。自動要約アルゴリズムを用いて、この各ドキュメントセットに対して、要約を作成する（ステップ１３）。ユーザーが、一つもしくはそれ以上の要約を選択することができ（ステップ１５）、選択した要約に対応するドキュメントセットが新たなドキュメントコレクションを形成する（ステップ２０）。ユーザーの満足するドキュメントセットを見つかるまで、このドキュメントコレクションの分割、要約及び選択の処理を繰り返して行う。



1

【特許請求の範囲】

【請求項１】　ユーザが探索するためのワードを入力することなくユーザが満足できるドキュメントのセットをドキュメントコレクションの中から探索するドキュメントコレクションの探索方法において、

（ａ）上記ドキュメントコレクションをドキュメントセットに分割し、

（ｂ）上記分割された各ドキュメントセットに対して各要約を作成し、

（ｃ）探索対象の上記ドキュメントコレクションを限定するために、少なくとも１つの上記要約から探索したい上記ドキュメントセットを選択して、これら選択されたドキュメントセットからなる新たなドキュメントコレクションを定義し、

（ｄ）ユーザが満足できるドキュメントセットが作られるまで上記（ａ）～（ｃ）を繰り返すことからなる、ドキュメントコレクションの探索方法。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は、ドキュメントコレクションの探索方法に関し、特にドキュメントの探索、分類及び探索の改良に関する。

【０００２】

【従来の技術】ますます多量のドキュメントが出版され、参考文献として利用できる現在の社会において、ドキュメントの探索あるいは検索は重要になってきている。このように膨大なドキュメントの中から、ユーザーの望む特別なドキュメントを探し出すことが難しいことが、膨大なドキュメントの抱える問題の一つである。

【０００３】ユーザーがキーワードあるいはフレーズを、例えば、コンピュータに入力すると、入力したキーワードあるいはフレーズを含むドキュメントを、ドキュメントの全体（あるいはこのドキュメントの全体から作られるワードインデクスあるいはルックアップテーブル）の中から探索できる多くのシステムが提案され、また今日稼動している。

【０００４】

【発明が解決しようとする課題】しかしながら、ユーザーが望むドキュメントあるいはドキュメントのセットにおいて使用されるワードあるいはフレーズが独特のものでない限り、数多くのドキュメントが探索され、扱いやすいヒット数に減らすためにユーザーに追加の入力を要求することがよくある。

【０００５】しかも、記事やドキュメントの著者によっては、異なるワードを同じか似た意味に用いることがよくある。ユーザーがさまざまな異なるワードを指定しない限り、入力されたワードあるいはフレーズによっては、関連するドキュメントが探索されないということがよくあることである。そこで、本発明の目的は、ユーザーが探索ワードを入力せずに、ドキュメントのコレクションの中から、ユーザーにとって満足のいくドキュメントセットを探索するドキュメント探索方法を提供することである。

2

【０００６】本発明の他の目的は、ドキュメント自動分割及びドキュメント自動要約アルゴリズムを用いて、ユーザーにとって満足のいくドキュメントセットを探索するドキュメント探索方法を提供することである。

【０００７】

【課題を解決するための手段】上記目的を達成するために本発明に係わるドキュメント探索方法は、ユーザが探索するためのワードを入力することなくユーザが満足できるドキュメントのセットをドキュメントコレクションの中から探索するものであり、（ａ）上記ドキュメントコレクションをドキュメントセットに分割し、（ｂ）上記分割された各ドキュメントセットに対して各要約を作成し、（ｃ）探索対象の上記ドキュメントコレクションを限定するために、少なくとも１つの上記要約から探索したい上記ドキュメントセットを選択して、これら選択されたドキュメントセットからなる新たなドキュメントコレクションを定義し、（ｄ）ユーザが満足できるドキュメントセットが作られるまで上記（ａ）～（ｃ）を繰り返すことからなる。

【０００８】上記ドキュメントコレクションのドキュメントセットへの分割及びドキュメントセットの要約作成ステップは、自動ドキュメント分割・要約アルゴリズムを実施することによって行う。このドキュメント自動分割アルゴリズムはクラスターアルゴリズムあるいはフラクションアルコリズムのような線型分割アルゴリズムを使用することが出来る。

【０００９】

【作用】本発明では、ドキュメントコレクションをドキュメントセットに分割して、この分割された各ドキュメントセットに対して各要約を作成して、少なくとも１つのこの要約から探索したい上記ドキュメントセットを選択することによって、ドキュメントコレクションを限定できる。これら選択されたドキュメントセットに対して、上記ドキュメントに分割、要約作成、ドキュメントセットの選択を繰り返すことによって、ユーザが探索するためのワードを入力することなくユーザが満足できるドキュメントのセットを探索することが出来る。

【００１０】

【実施例】以下、図面を参照して、本発明の実施例を説明する。本発明では、ドキュメントのコレクションの中から、ユーザーが望むドキュメントあるいはドキュメントセットを識別するためコンピュータを援用する。図１は、本発明に係わるドキュメント探索方法を示すフローチャートである。以下、図１に基づいてドキュメント探索方法を説明する。ステップ１１において、ドキュメントコレクションをセットに分割するプログラムあるいはアルゴリズムを用いて、ドキュメントコレクションをセ

ットに分割する。このプログラムは、例えば、ワードの頻度、キーワードの存在あるいは他の基準を用いた所定の基準に従って、自動的にドキュメントを分割するプログラムを使用することが出来る。分割アルゴリズムは、例えば、公知であるフラクションあるいはクラスタアルゴリリズムを使用することが出来る。分割プログラムは、探索対象のドキュメントコレクションの含まれるドキュメントの数が膨大になれば、特に線型的であることが望ましいが、ドキュメントの数に対して、幾何級数的あるいは指数的に増加するものであっても使用可能である。

【００１１】使用可能なクラスタアルゴリズムについての論文が、Ｊonesによる１９９１年２月発行の”Notes and reference on early automatic classification work”の10～17ページに記載されており、参考のために引用した。使用できるアルゴリズムの一つの例として、１９８５年ACM の197 ～203 ページ、Yu他による”Adaptive Document Clustering”記載されており、参考のために引用した。階層化ドキュメントクラスタ化の別の一例が、１９８８年のInformation Processing & Management の vol.24 、No.5、577 ～597 ページ、Willett による”Recent Trends in Hierarchic Document Clustering:A Critical Review ”に記載されており、参考のために引用した。

【００１２】他にも公知であるドキュメント分割技術があり、これらの技術を使用することによってドキュメントコレクションをセットに分割することが出来る。いくら基準の数があっても、この基準に従って、セットを決定することが出来る。例えば、上述したInformation Processing & Management のvol.24、No.5、577 ～597 ページ、”Recent Trends in Hierarchic Document Clustering:A Critical Review ”において、Peter Willett は、ドキュメント探索のための階層型凝集クラスタ化方法の使用について論じている。Willett はそこで、ドキュメント間の類似についての計算及びドキュメントのクラスタ化にとって適切なクラスタ化方法を導入し、これらの方法を平凡でない大きさのデータベースに実施できるようなアルゴリズムについて論じており、ランダムグラフ理論とクラスタ化すべきドキュメントコレクションの経験的な特徴に基づたテストを行って、ドキュメントの階層化ができることを確認している。Willett は、階層化されたドキュメントを探索範囲とすることができることを示している。

【００１３】また、数種類の異なるタイプの階層型凝集クラスタ化方法を用いて、クラスタ化を行ない、その結果生じたクラスタを探索すべき範囲として使用した一連の研究プロジェクトの結果が示されている。また、完全なリンケージ方法(complete linkage method) 、最も近接する近接クラスタ方法(nearest neighbor clustermethod) 等が論じられている。

【００１４】また、ドキュメントの内容を識別する他の例として、１９９０年６月にニュージャージ州のアトランタでの第１０回パターン認識国際会議において、Tsujimoto 他によって発表された、”Understanding Multi-Articled Document ”がある。この論文は、ドキュメントを文字認識してこれによって内容の意味を判別するというやり方によらず、ドキュメントを理解する方法を論じている。それは、ドキュメントが明らかに幾何的な階層構造をもっており、少しの規則を用いるだけで、この幾何的な階層構造をドキュメントのもつ意味を表す論理構造に変換できることを示している。

【００１５】上述したドキュメントコレクションの分割によって生じるセットの数は所望により調整することが出来るが、ユーザーに関心のありそうなドキュンメントを効率的に選択したり、効率的に分離するためにユーザーが簡単にソートできる程度の数が望ましい。セットの数は１５～２０の数が理想的である。ドキュメントがセットに分割されたあと、ステップ１３において分割された各セットに対して要約が作成される。この要約は、例えば、公知の自動要約アルゴリズムを用いて作成される。

【００１６】図２は、要約処理を示すフローチャートである。ステップ３１において、ドキュメントのなかでその内容を良く表すワードが決定され、ステップ３２においてそのワードの使用頻度が決定される。ドキュメントのなかでその内容を良く表すワードは、例えば、各ドキュメントのなかで最も使用頻度の高いワードを見つけだす逆ドキュメント使用頻度法（Inverse Document Frequency(IDF) ）を用いて、識別される。ステップ３２において、そのワードを含むドキュメント内の文章が表示される。

【００１７】要約が、例えば、コンピュータモニタ上に表示される。図１のステップ１５において、表示された要約に対して、例えば、マウス、ジョイスチック等のコンピュータへの入力手段を用いて、１つもしくはそれ以上の要約を選択することによって、対応した１つもしくはそれ以上のドキュメントのセットを選択することが出来る。

【００１８】ステップ１５において、ユーザーは表示されたドキュメントセットが満足するものであるかどうかを判断する。そのドキュメントセットが満足できる場合は、ステップ１８において、そのセットに含まれるドキュメントの内容が、印刷あるいは、熟読等のために、ユーザーに表示される。一方、選択されたセットに余り多くのドキュメントが含まれているようだと、選択されるドキュメントの数を減らすために、以下の処理がされる。

【００１９】ステップ２０において、ユーザーによってドキュメントの分割によって作成されたドキュメントの各セットに対して一つあるいはそれ以上の選択されたセ

5

ットがドキュメントの新しいコレクションを形成する。この新しいドキュメントのコレクションが、ステップ１３において、再度分割され、ステップ１５において、再度要約が作成される。そして、この再度作成された要約が、ユーザーに表示され、希望する場合は更に再選択される。この時、分割アルゴリズムは、選択されたドキュメントセットのなかでドキュメントの差に基づいてさらに細かいパーティションを作成するタイプであり、ドキュメントが前に分割されたコレクションにおいて、全ドキュメントに共通な特別な探索ワード基づいて、分割を行うものでないことが望ましい。

【００２０】このような処理を繰り返し行なって、ユーザーにとって満足できる最終的なセットに分割されるまで、ドキュメントのセットが小さくされる。このように、分割、要約、セットの選択の各ステップを繰り返し

6

行うことによって最終的にユーザーの望むセットにドキュメントを分割・選択ができる。発明は、ある程度特別な場合において説明したが、勿論これに限定されることなく、その組合せ等を変更して使用することが可能である。

【００２１】

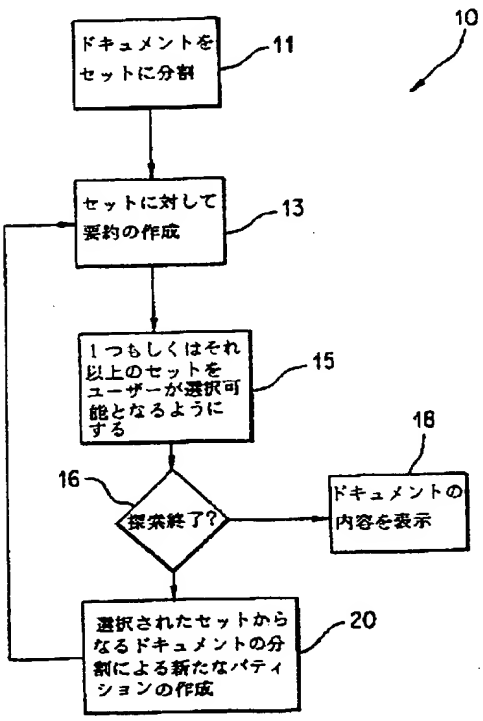
【発明の効果】以上説明したように本発明では、ユーザーが探索ワードを入力せずに、ドキュメントのコレクションの中から、ユーザーにとって満足のいくドキュメントセットを探索することができる。

【図面の簡単な説明】

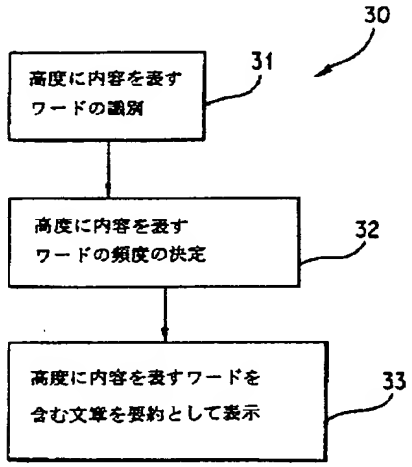
【図１】ドキュメント探索方法を示すフローチャートである。

【図２】要約処理を示すフローチャートである。

【図１】



【図２】



フロントページの続き

(72)発明者　マイケル　ジェイ．バーバリーノ
アメリカ合衆国　カリフォルニア州
94038　モス　ビーチ　ピー．オー．ボックス　853